2020

Mock Jurors' Evaluation of Firearm Examiner Testimony

Brandon L. Garrett
Duke University

Nicholas Scurich
Duke University

William E. Crozier
Duke University

Mock Jurors' Evaluation of Firearm Examiner Testimony

Abstract

Objectives: Firearms experts traditionally have testified that a weapon leaves "unique" toolmarks, so bullets or cartridge casings can be visually examined and conclusively matched to a particular firearm. Recently, due to scientific critiques, Department of Justice policy, and judges' rulings, firearms experts have tempered their conclusions. In two experiments, we tested whether this ostensibly more cautious language has its intended effect on jurors (Experiment 1), and whether cross-examination impacts jurors' perception of firearm testimony (Experiment 2). Hypotheses: Four hypotheses were tested. First, jurors will accord significant weight to firearm testimony that declares a "match" compared to testimony that does not (Experiments 1 and 2). Second, variations to "match" language will not affect guilty verdicts (Experiment 1). Third, only the most cautious language ("cannot exclude the gun") would lower guilty verdicts (Experiment 1). Fourth, cross-examination will reduce guilty verdicts depending on specific language used (Experiment 2). Method: In two preregistered, high-powered experiments with 200 mock jurors per cell, participants recruited from Qualtrics Panels were presented with a criminal case containing firearms evidence, which varied the wording of the examiner's conclusion and whether cross-examination was present. These variations include conclusion language used by practitioners, language advised by government organizations, and language required by judges in several cases. Participants gave a verdict, rated the evidence and expert in all conditions. Results: Guilty verdicts significantly increased when a match was declared compared to when a match was not declared. Variation in conclusion language did not affect guilty verdicts nor did it affect jurors' estimates of the likelihood the defendant's gun fired the bullet recovered at the crime scene. In contrast, however, a more cautious conclusion that an examiner "cannot exclude the defendant's gun" did significantly reduce guilty verdicts and likelihood estimates alike. The presence of cross-examination did not affect these findings. Conclusion: Apart from the most limited language ("cannot exclude the defendant's gun"), judicial intervention to limit firearms conclusion language is not likely to produce its intended effect. Moreover, cross-examination does not appear to affect perceptions or individual juror verdicts.

Disciplines

Forensic Science and Technology

Comments

This article is published as Garrett, B. L., Scurich, N., & Crozier, W. E. (2020). Mock jurors' evaluation of firearm examiner testimony. *Law and Human Behavior, 44*(5), 412–423. https://doi.org/10.1037/lhb0000423. Posted with permission of CSAFE.

Mock Jurors' Evaluation of Firearm Examiner Testimony

Brandon L. Garrett
Duke University

Nicholas Scurich
University of California, Irvine

William E. Crozier
Duke University

Objectives: Firearms experts traditionally have testified that a weapon leaves "unique" toolmarks, so bullets or cartridge casings can be visually examined and conclusively matched to a particular firearm. Recently, due to scientific critiques, Department of Justice policy, and judges' rulings, firearms experts have tempered their conclusions. In two experiments, we tested whether this ostensibly more cautious language has its intended effect on jurors (Experiment 1), and whether cross-examination impacts jurors' perception of firearm testimony (Experiment 2). ***Hypotheses:*** Four hypotheses were tested. First, jurors will accord significant weight to firearm testimony that declares a "match" compared to testimony that does not (Experiments 1 and 2). Second, variations to "match" language will not affect guilty verdicts (Experiment 1). Third, only the most cautious language ("cannot exclude the gun") would lower guilty verdicts (Experiment 1). Fourth, cross-examination will reduce guilty verdicts depending on specific language used (Experiment 2). ***Method:*** In two preregistered, high-powered experiments with 200 mock jurors per cell, participants recruited from Qualtrics Panels were presented with a criminal case containing firearms evidence, which varied the wording of the examiner's conclusion and whether cross-examination was present. These variations include conclusion language used by practitioners, language advised by government organizations, and language required by judges in several cases. Participants gave a verdict, rated the evidence and expert in all conditions. ***Results:*** Guilty verdicts significantly increased when a match was declared compared to when a match was not declared. Variation in conclusion language did not affect guilty verdicts nor did it affect jurors' estimates of the likelihood the defendant's gun fired the bullet recovered at the crime scene. In contrast, however, a more cautious conclusion that an examiner "cannot exclude the defendant's gun" did significantly reduce guilty verdicts and likelihood estimates alike. The presence of cross-examination did not affect these findings. ***Conclusion:*** Apart from the most limited language ("cannot exclude the defendant's gun"), judicial intervention to limit firearms conclusion language is not likely to produce its intended effect. Moreover, cross-examination does not appear to affect perceptions or individual juror verdicts.

Public Significance Statement
This study addresses mounting legal concerns regarding the overstated conclusions experts reach using firearms comparison, one of the most commonly used forensic disciplines in criminal cases.

Bradley D. McAuliff served as Action Editor.

Brandon L. Garrett, School of Law, Duke University; Nicholas Scurich, Department of Psychological Science and Criminology, Law and Society, University of California, Irvine; William E. Crozier, School of Law, Duke University.

This work was funded by the Center for Statistics and Applications in Forensic Evidence through Cooperative Agreement 70NANB20H019 between the National Institute of Standards and Technology and Iowa State University, which includes activities carried out at Carnegie Mellon University; Duke University; University of California, Irvine; University of Virginia; West Virginia University; University of Pennsylvania; Swarthmore College; and University of Nebraska, Lincoln. The authors have no conflicts of interest to report. All materials, data, analyses, results, and preregistration are available on the Open Science Framework (Study 1: https://osf.io/7kb43/; Study 2: https://osf.io/gkjts/). The findings in this article have not been previously presented.

The data are available at https://osf.io/qfsc3/?view_only=5f34d29b870b4adeb9f15d87dc23392a

The experiment materials are available at https://osf.io/qfsc3/?view_only=5f34d29b870b4adeb9f15d87dc23392a

The preregistered design and analysis plan is accessible at https://osf.io/qfsc3/?view_only=5f34d29b870b4adeb9f15d87dc23392a

Correspondence concerning this article should be addressed to Nicholas Scurich, Department of Psychological Science and Criminology, Law and Society, University of California, Irvine, 4213 Social and Behavioral Sciences Gateway, Irvine, CA 92697-7085. E-mail: nscurich@uci.edu

While judges and prosecutors have required experts to use more modest phrasing in their conclusions, the effectiveness of those changes was unknown. This study found such changes largely ineffective, suggesting more aggressive regulation is needed to prevent jurors and legal actors from misunderstanding firearms evidence.

Keywords: forensic science, expert testimony, error rates, juror decision-making, firearms

Supplemental materials: http://dx.doi.org/10.1037/lhb0000423.supp

Firearms violence is a sizable problem in the United States, with over 10,000 homicides involving firearms and almost 500,000 other crimes, such as robberies and assaults, committed using firearms each year (Federal Bureau of Investigation, 2018; Bureau of Justice Statistics, 2011). For that reason, firearms comparisons are in great demand as one of the most commonly performed forensic analysis. Such forensic comparisons seek to potentially link crime scene evidence, such as spent cartridge casings or bullets, with a particular firearm. For over a hundred years, firearms experts have testified in criminal trials in the United States (Lanigan, 2012). Firearms experts traditionally testify in court that a weapon will leave toolmarks on ammunition fired, so that bullets or cartridge casings can be visually examined and then conclusively matched to a firearm. By the late 1990s, such testimony was premised on a "theory of identification" promulgated by a professional association, the Association of Firearms and Tool Mark Examiners (AFTE). AFTE currently instructs practitioners to use the term "identification" to explain what it means when they identify "sufficient agreement" when comparing cartridge casings or bullets, based on the "examiner's training and experience" (AFTE, 1998, 2020).

Criticism of forensic science by academics is not new (Kennedy, 2003; see generally Mnookin, Cole, Dror, & Fisher, 2010). In recent years, however, authoritative scientific organizations have called into question the validity and the reliability of such categorical testimony regarding firearms comparison evidence. In a 2008 report on ballistic imaging, the National Academy of Sciences concluded that definitive associations of that type were not supported (National Research Council, 2008). In a 2009 report covering a broad range of forensic disciplines, the National Academy of Sciences stated that such categorical conclusions regarding firearms or toolmarks were not supported by research, and that instead, more cautious comparative claims should be made (National Research Council, 2008). The report stated that the "scientific knowledge base for tool mark and firearms analysis is fairly limited" (p. 155). The report noted that an examiner makes "a subjective decision based on unarticulated standards and no statistical foundation for estimation of error rates" (pp. 153–154). The National Academy of Sciences report added that the AFTE theory of identification is inadequate and does not explain how an expert can reach a given level of confidence in a conclusion. Going farther, the report of the President's Council of Advisors on Science and Technology (2016), after reviewing the extant studies of firearm examiner performance, found firearms comparison methods not foundationally valid.

Anticipating some of these concerns, several judges have intervened to limit the scope of expert testimony concerning firearms, although no judge in the United States has excluded firearms testimony altogether. For example, judges have required that examiners opine only that it is "more likely than not" that the ammunition could have come from the defendant's firearm (*United States v. Glynn,* 2008). Others have required that the examiner limit conclusions to a "reasonable degree of ballistic certainty" (*United States v. Monteiro,* 2006; *United States v. Diaz,* 2007). Another judge limited testimony to observing that the markings were "consistent" (*United States v. Willock*, 2010).

In January 2019, the U.S. Department of Justice (DOJ) responded to these concerns and announced new guidelines for firearms testimony, stating that firearms experts cannot testify in federal case that toolmarks originated from the same source, to the exclusion of all others. Instead, an expert should testify as to a "source identification." That is defined as follows:

> "Source identification" is an examiner's conclusion that two toolmarks originated from the same source. This conclusion is an examiner's decision that all observed class characteristics are in agreement and the quality and quantity of corresponding individual characteristics is such that the examiner would not expect to find that same combination of individual characteristics repeated in another source and has found insufficient disagreement of individual characteristics to conclude they originated from different sources. (U. S. DOJ, 2019, p. 2)

In one of the most far-reaching judicial rulings to date, the firearm expert was restricted to a conclusion that he "cannot exclude" the relevant firearm as the source of the fired cartridge casings (*United States v. Tibbs,* 2019). The judge in that case was concerned that the evidence not be barred: "I want to make clear I am not excluding specialized opinion testimony in the area of firearms and toolmark examination as a whole, or finding that the entire discipline lacks foundational reliability" (*United States v. Tibbs,* 2019, p. 57). The judge explained in more detail what this "cannot exclude" restriction means in practice for the expert:

> He [the firearm examiner] will be able to describe the work he did and the comparisons he made. He can make, as the defense concedes, a comparison based on class characteristics and he can conclude that based on his examination the recovered firearm cannot be excluded as the source of the shell casings found on the scene of the alleged shooting.
>
> He cannot state an ultimate conclusion in different or stronger terms, and he cannot state that individual marks are unique to a particular firearm such that an identification could be made. (*United States v. Tibbs,* 2019, p. 58)

It is unknown, however, whether these efforts by judges and federal prosecutors to limit firearms testimony affects how jurors

evaluate such testimony. Research has been done on a range of other forensic techniques, including DNA testimony involving statistical methods, as well as traditional forensics like latent fingerprint comparisons in which experts traditionally offer "identification" conclusions (Schklar & Diamond, 1999; Scurich & John, 2013; Thompson, Kaasa, & Peterson, 2013). As a general matter, the laity views forensic science as highly accurate and persuasive (see Lieberman, Carrell, Miethe, & Krauss, 2008), and is generally insensitive to variations in the way in which a "match" is communicated in non-numeric terms (Thompson & Newman, 2015; McQuiston-Surrett & Saks, 2009). For example, prior research has found that jurors place great weight on fingerprint evidence and regard it as accurate and reliable, regardless of whether the expert expresses conclusions in more certain or more cautious terms, such as a simple match, a match individualized to the defendant, a match made to a scientific certainty, or a match individualized to the defendant that is practically impossible to have come from any other person (Garrett & Mitchell, 2013; Garrett, Mitchell, & Scurich, 2018). However, there is some evidence to suggest that the weight mock jurors place on forensic evidence varies depending on the forensic discipline (Garrett, Crozier, & Grady, 2020; Ribeiro, Tangen, & McKimmie, 2019). Thus, it is important to understand how laypeople weigh non-numeric conclusions for firearms testimony.

Study 1

Despite its common use and importance in criminal trials, however, little empirical research has been done to on the impact of firearms testimony on jurors. In fact, the only empirical study on the perception of firearm examiners was conducted by Saks and Wissler in 1984. That study simply asked participants ($n = 97$) to rate the competence and likelihood they would agree with different types of experts (ratings were made on a 1–10 scale with 10 indicating the highest competence/likelihood of agreement); firearm examiners received the third highest overall ratings (mean competence rating = 7.80; mean likelihood of agreement rating = 7.62). Only medical doctors and chemical/drug experts received higher ratings.

The study by Saks and Wissler (1984) is consistent with the notion that jurors perceive firearm testimony to be powerful evidence. However, the study leaves open many unanswered questions. First, would the gist of the findings replicate 35 years later on a broader sample of adults? It seems plausible that views about forensic science may have changed in the interim, perhaps due to changes in the manner in which forensic science is portrayed in the media (see Tyler, 2005). Second, although it is clear that participants in the study indicated that they are likely to agree with the testimony of a firearm examiner in general, would jurors agree with any particular firearm examiner testimony in the context of a criminal trial? This distinction between anticipated and actual behavior could tested by viewing the behavior of jurors after receiving firearm testimony or not. Finally, would variations in the language used by a firearm examiner to characterize a "match" impact how jurors evaluate firearm evidence and in turn influence their verdicts? This question is particularly germane in light of the recent reforms to curtail the language used by firearm examiners in actual court proceedings.

The present studies seek to address these lacunae. Specifically, in this study a large, nationally representative sample of mock jurors was presented with a criminal case in which a firearm expert drew a conclusion regarding his analysis of the casing recovered from the crime scene and the defendant's gun. In Study 1, the language he used to characterize the match was experimentally manipulated. Participants were then furnished with judicial instructions on the standard of proof, and asked them to render a verdict on whether to convict and the defendant or not, rate the quality of the evidence, the expert, and other aspects of the case. Study 2 first attempted to replicate the principal findings in Study 1 on the effect of different language, and also tested what impact if any cross-examination might have—either alone or in combination (i.e., an interaction) with the variations in language used to describe the match.

Hypotheses

Three main hypotheses are put forth for Study 1:

Hypothesis 1: Consistent with how the laity views forensic science in general (Lieberman, Carrell, Miethe, & Krauss, 2008), and firearm experts in particular (Saks & Wissler, 1984), we hypothesized that jurors will accord significant weight to firearm testimony that declares a "match" between two cartridge casings. That is, guilty verdicts, and the likelihood of the defendant's guilt, would be higher for conclusions that match the evidence casing to the defendant's gun, compared to a condition that does not conclude a match (i.e., an Inconclusive result).

Hypothesis 2: Consistent with previous research finding that variations to non-numeric language used to characterize a "match" does not have an impact on juror's evaluation of fingerprint evidence (Garrett & Mitchell, 2013), we hypothesized that variations to the language used by firearm examiners to characterize a match would not affect guilty verdicts.

Hypothesis 3: We hypothesized that only the most cautious language (i.e., "cannot exclude the gun") would lower guilty verdicts compared to other conclusion language.

The second hypothesis directly tests assumptions made by judges and prosecutors that more limited, less certain language will have a corresponding effect on jurors—the ultimate consumers of firearm examiner testimony. We test several different variations to the language, all of which come from case law or official organizations, such as DOJ, that have authority and actually impact the manner in which testimony is presented in court. In short, these language variations have a high degree of ecological validity.

It is important to note that Hypothesis 2 predicts a null effect. This prediction has its grounding in studies that also found a null effect of "match" language used for other forms of forensic science evidence (e.g., Garrett & Mitchell, 2013). Given that a null hypothesis is predicted, it is extremely important to have sufficient statistical power to detect an effect should such an effect actually exist. An underpowered study might fail to reject the null hypothesis—thus confirming our prediction of a null effect—simply because it lacks the power to detect an effect. To minimize this possibility, we opted to have a high-powered study that should be

able to detect even small effects (assuming they do in fact exist). As such, we elected to have 200 participants per cell, which according to statistical power calculations is sufficient to detect a small effect ($d = 0.35$) with power = 0.90 and $\alpha = .05$ using a two-tailed t test for any differences between groups.

Method

Participants. One thousand four hundred twenty participants were recruited using Qualtrics and completed the survey online. Participants were terminated from the study and excluded from analyses if they did not pass an instructional manipulation check to assess attention or reading checks, if they were not a U.S. citizen, or if they came from suspicious, duplicate geolocations (see Oppenheimer, Meyvis, & Davidenko, 2009). The final sample was comprised of 1420 participants aged 18–88 ($Mdn = 47$, interquartile range = 28). The study was balanced with respect to gender (48% self-identified as male; 52% as female). Participants self-identified as 13% Black, 5% Asian, 60% White, 17% Hispanic, 1% Native American or Pacific Islander, and the rest selected other. With respect to education, 65% had a 2-year college degree or less, 25% had at least 4-year college degree, and 10% had a postgraduate degree.

Participants were asked to self-identify their political preferences: 10.8% identified as "very conservative," 18.1% identified as "somewhat conservative," 39.4% identified as "middle of the road," 19.2% identified as "somewhat liberal," and 12.5 identified as "very liberal." Nineteen percent had a self-reported annual household income of less than $20,000 and 15% had an income above $100,000. Thirty-seven percent of the participants reported having served previously on a jury.

Materials.

The case. Each participant read a synopsis of a criminal case against a defendant for discharging a firearm in a public place. The facts in the synopsis were adapted from an actual criminal case (*United States v. Driscoll*, 2003). Briefly, the police arrested the defendant ("Mr. Cole") for firing a gun in an unsuccessful convenience store robbery. The defendant in the case was charged with willfully firing a firearm during the commission of a felony, and attempted armed robbery. During the attempted convenience store robbery, no one was hurt as the gun was fired into the floor; the culprit fled and the clerk could not identify any person, as the gunman wore a mask. However, two days later, during a routine traffic stop, police pulled over the defendant and confiscated his 9-mm handgun, as a 9-mm bullet had been found at the crime scene. The case materials did not include a cross-examination, opening or closing arguments, or any witnesses or evidence beside the firearm expert (see below). The length of the stimulus was approximately 1,300 words. Full text can be found in the online supplemental materials.

Direct examination of the firearm expert. In a question-and-answer transcript format, a firearm expert detailed his experience, firearm examination methodology, and his analysis in this case. He then reached one of the seven possible conclusions:

1. *Inconclusive:* The examiner testified that "there was not sufficient agreement among the characteristics to determine that the toolmarks have been produced by the same tool."
2. *Cannot be excluded:* The examiner testified "the defendant's gun cannot be excluded as the gun that fired the bullet recovered at the crime scene," which means that "we observe some agreement of a combination of individual characteristics suggesting that the toolmarks have been produced by the same tool."
3. *Simple identification:* The examiner testified that "I identified the crime scene bullet as having been fired by the defendant's gun."
4. *Ballistic certainty:* The examiner testified that "I identified the crime scene bullet as having been fired by the defendant's gun," and that this "means that the bullet recovered from the crime scene was identified, to a reasonable degree of ballistic certainty, as having come from the defendant's firearm (*United States v. Diaz,* 2007).
5. *More likely than not:* The examiner testified that "I identified the crime scene bullet as having been fired by the defendant's gun" and that this "means that it is more likely than not that the bullet recovered from the crime scene came from the defendant's firearm" (*United States v. Glynn, 2008*).
6. *Complete agreement:* The examiner testified that "I concluded that the crime scene bullet has markings consistent with being fired by the defendant's gun," and that this "means that when I looked at the class characteristics of the bullet recovered from the crime scene and a bullet fired through the defendant's gun, that the class characteristics were in complete agreement" (*United States v. Monteiro, 2006*).
7. *DOJ:* The examiner testified that "I identified the crime scene bullet as having been fired by the defendant's gun," and this "means that the bullet recovered at the crime scene and a bullet fired through the defendant's gun originated from the same source. That all the class characteristics are in agreement and I would not expect to find that same combination of agreement in another source."

Dependent measures. After the direct examination of the firearm expert, participants were provided instruction on the standard of proof (i.e., beyond a reasonable doubt) asked whether, given the state's burden to prove guilt beyond a reasonable doubt, whether they would convict the defendant (dichotomous "guilty" or "not guilty") and the likelihood that the defendant was the man who fired the gun (0–100%). Participants then responded to six different items that directly probed the credibility and reliability of the firearm analysis presented in the case (e.g., "How reliable do you think the firearm evidences is in this case?") and in general (e.g., "In general, how often do firearm examiners make mistakes when determining whether bullets were fired through the same gun?"). These items were rated on a 7-point Likert scale with higher values indicating high reliability/credibility/validity. A scale analysis revealed a high degree of correlation among the six items, yielding a Cronbach's $\alpha = .88$. Thus, a composite score (referred to as

"scientific credibility") was created by summing the scores and dividing by 6. Finally, participants were asked general questions about their views about firearms evidence ("How often do firearm examiners make mistakes?" 1 = *never* to 7 = *always*; "Do you think firearm evidence is generally reliable [scientific]?" 1 = *strongly agree* to 7 = *strongly disagree*; "Do you think guns leave unique markings on discharged bullets/casings?" *yes*/*I do not know*/*no*), as well as whether erroneous convictions or failing to convict a guilty person cause more harm to society.

Finally, we included two additional questions: a 1–9 rating of the strength of the case against the defendant; and "Which of the following errors at trial do you believe causes more harm to society" with "failing to convict a guilty person," erroneously convicting an innocent person" or "both are equally bad" as responses. These questions were exploratory and we did not analyze them because the other dependent measures sufficiently addressed our research questions. All of the study materials, as well as underlying data, are available on an Open Science Framework website: https://osf.io/7kb43/.

Procedure. This study was approved by the Duke University Institutional Review Board. The participants first provided informed consent and responded to basic eligibility questions such as age and U.S. citizenship, as well as questions regarding demographics, income, and political preferences. Each participant then read the crime scenario describing a simple criminal offense and police investigation. Participants were then randomly assigned to one of the seven conclusion conditions described above. Finally, participants completed the dependent measures and were thanked for their participation. Average completion time was 780.61 s (SD = 1,862.88).

Results

At the conclusion of trial, 50.4% (n = 715) of participants voted to convict the defendant. The percentage of guilty verdicts within each experimental condition appears in Figure 1.

Three separate binary logistic regressions were conducted to test whether the proportion of guilty verdicts varied as a function of the experimental conditions. All models use a bootstrap procedure with 1,000 resamples to estimate standard errors of measurement. Model 1 uses the "inconclusive" condition as the referent category, Model 2 uses the "cannot be excluded" group as the referent category, and Model 3 uses the "simple identification" group as the referent category. Note that Model 1 tests Hypothesis 1, Model 2 tests Hypothesis 3, and Model 3 tests Hypotheses 2. The results of all models are presented in Table 1.

With regard to Model 1, the model fit was statistically significant (χ^2 = 118.84, df = 6, $p < .001$, Nagelkerke R^2 = 0.11), and every experimental condition was significantly more likely to convict the defendant than the "inconclusive" condition. Indeed, even the "cannot be excluded" condition which had the smallest relative effect size, increased the odds of conviction by 2.5 (Exp[B] = 2.497, 95% confidence interval [CI] [1.6, 3.9]) compared to the inconclusive condition. Regardless of the specific language used, the odds of conviction increased by 5 to 6 when a match was declared relative to when an inconclusive was declared. For example, the odds of conviction increase by 6 (Exp[B] = 6.2, 95% CI [4.0, 9.6]) when a "simple match" is declared compared to an inconclusive. This finding corroborates Hypothesis 1 that firearm examiner testimony constitutes powerful evidence.

With regard to Model 2, the model fit was statistically significant (χ^2 = 29.1, df = 5, $p < .001$, Nagelkerke R^2 = 0.32), and every experimental condition was significantly more likely to convict the defendant than the "cannot be excluded" condition. Participants in the "simple identification" condition were over two times more likely to convict the defendant (Exp[B] = 2.41, 95% CI [1.6, 3.6]) than participants in the "cannot be excluded" condition. Similarly, participants in the "ballistic certainty" condition were 1.9 (95% CI [1.3, 2.8]) times more likely to convict, participants in the "more likely than not" condition were 2.3 (95% CI [1.6, 3.4]) times more likely to convict, participants in the "com-

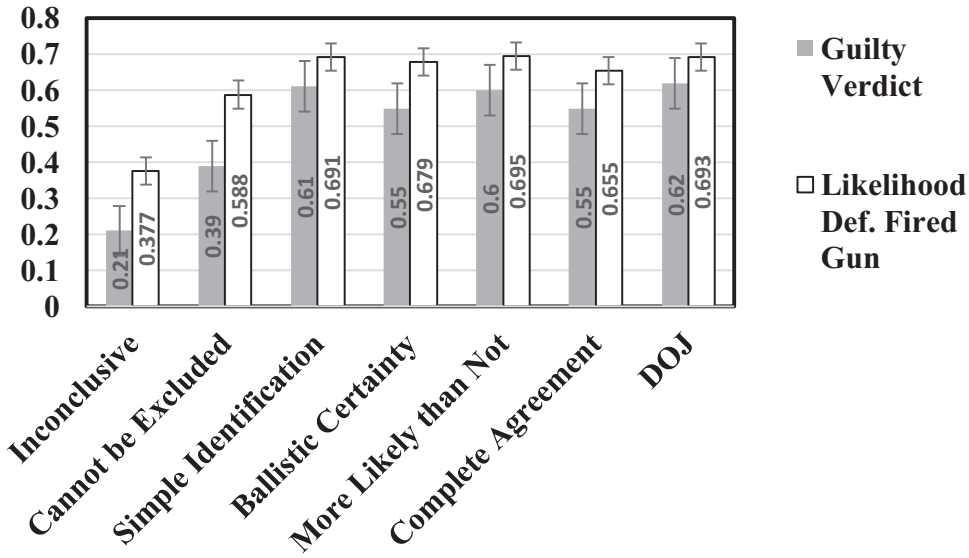

Figure 1. Proportion of guilty verdicts and mean likelihood ratings that the defendant fired the gun (with 95% confidence intervals) in each experimental condition. DOJ = Department of Justice.

Table 1
Study 1 Binary Logistic Regression Predicting Verdict (1 = Guilty, 0 = Not Guilty)

Experimental condition	Model 1	Model 2	Model 3
Constant	−1.35 [.001] (0.17)	−0.43 [.002] (0.14)	−0.45 [.003] (0.15)
Cannot be excluded	0.92 [.001] (0.22)		
Simple identification	1.79 [.001] (0.23)	0.88 [.001] (0.21)	
Ballistic certainty	1.56 [.001] (0.22)	0.65 [.001] (0.20)	−0.23 [.26] (0.21)
More likely than not	1.75 [.001] (0.223)	0.84 [.002] (0.202)	−0.040 [.85] (0.207)
Complete agreement	1.53 [.001] (0.223)	0.61 [.002] (0.201)	−0.267 [.20] (0.206)
Department of Justice	1.82 [.001] (0.22)	0.91 [.001] (0.20)	0.300 [.89] (0.21)
N	1,415	1,207	999
Nagelkerke R^2	0.11	0.32	0.01

Note. Estimated raw maximum likelihood binary logistic regression weights, with standard errors in parentheses and *p* values in brackets.

plete agreement" condition were 1.8 (95% CI [1.2, 2.7]) times more likely to convict, and participants in the "DOJ" condition were 2.5 (95% CI [1.7, 3.7]) times more likely to convict than participants in the "cannot be excluded" condition. This finding—that the conviction rate is higher in all other groups than the "cannot be excluded group"—corroborates Hypothesis 3.

With regard to Model 3, which uses the "simple identification" as the referent category and omits the "inconclusive" and "cannot be excluded" conditions, the model fit is not statistically significant ($\chi^2 = 3.7$, $df = 4$, $p = .45$, Nagelkerke $R^2 = 0.01$). None of the experimental conditions differed statistically. In other words, variations to the language used to characterize the match did not affect verdicts beyond declaring a "simple match." This corroborates Hypothesis 2.

Participants' ratings of the likelihood that the defendant was the man who fired the gun (0–100%) was submitted to a one-way analysis of variance (ANOVA) with follow-up Bonferroni-adjusted contrasts. Note the cell means (and 95% CIs) are plotted in Figure 1. The ANOVA was significant, $F(6, 1408 = 36.78, p < .001$, and mirrored the results of the conviction rates. The likelihood of the defendant's guilt was lowest for the "inconclusive" condition ($M = 37.65$, $SD = 27.92$) which was significantly different from all other conditions ($ps < .001$, $d = [-1.1, -0.74]$). The second lowest-rated condition was the "cannot be excluded" condition ($M = 58.75$, $SD = 29.00$), which was significantly lower than the "simple identification," $t(397) = -3.78$, $p = .003$, $d = -0.37$, "ballistic certainty," $t(393) = -3.40$, $p = .014$, $d = -0.34$, "more likely than not," $t(392) = -3.95$, $p = .002$, $d = -0.37$, and "DOJ," $t(393) = -3.89$, $p = .002$, $d = -0.37$, but not significantly different from the "complete agreement" condition, $t(387) = -2.45$, $p = .30$, $d = -0.24$.

A one-way ANOVA with scientific credibility as the dependent variable was significant $F(6, 1408 = 2.32, p = .031)$, though post hoc Bonferroni contrasts revealed that the only statistically significant difference was that cannot be excluded ($M = 2.26$, $SD = 0.86$) was viewed as more scientifically credible than the more likely than not condition ($M = 2.0$, $SD = 0.90$; $p = .049$, 95% CI [.0003, .5319], $d = 0.30$). This finding indicates that scientific credibility is not dependent on the outcome (e.g., an inconclusive result is no less credible than an identification).

Finally, because no studies have used firearms and ballistic forensic evidence as stimuli, participants' general views of the forensic discipline were measured. The data revealed that participants, on average, believe firearm experts do not make mistakes in their analysis very often ($M = 2.66$, $SD = 0.88$, falling between *almost never* and *sometimes* anchors on a 1–7 Likert scale with lower scores indicating fewer mistakes). When asked to rate agreement with whether firearm evidence is, in general, reliable, participants on average agreed ($M = 2.25$, $SD = 1.24$, falling between *agree* and *somewhat agree* anchors on a 1–7 scale), and when asked to rate agreement with whether firearm evidence is, in general, scientific, participants also on average agreed ($M = 2.14$, $SD = 1.13$, falling between *agree* and *somewhat agree* anchors on a 1–7 scale). An overwhelming majority (84.5%) of participants ($n = 1200$) stated that they believe firearms leave unique markings on bullet casings; while only 12.9% ($n = 183$) said that they were unsure, and 2.6% ($n = 37$) said firearms do not leave unique markings.

Discussion

Despite how often firearm examiners testify in criminal trials, there has yet to be an empirical study of how jurors perceive firearm examiner testimony. In the first study to examine how mock jurors perceive and evaluate firearm forensic evidence, the data reveal that mock jurors accord significant weight to a firearm examiner declaring match. This effect is seen in both guilty verdicts and likelihood ratings that compare an inconclusive conclusion to a match—regardless of the language used to describe the match. Another key finding is that variations to the conclusion language used to describe a match (i.e., not an inconclusive finding) do not appear to matter to jurors, with the exception of the cannot exclude language. Jurors were just as likely to convict and gave the same likelihood ratings when the firearm examiner used "simple identification" language as the other formulations. Although this effect is consistent with other research in the domain of forensic fingerprint evidence (Garrett & Mitchell, 2013), it calls into question the approach of modifying the language used by firearm examiners in an effort to avoid overstating to jurors. These data indicate that tweaks to the conclusory language do not matter to jurors and produce no meaningful effect on a trial outcome. The exception to this observation is the cannot exclude language; this approach does appear to have an impact on jurors' perception of firearm evidence.

There is an important caveat to these findings. The study did not include a cross-examination condition, in which the expert is probed regarding his or her conclusions and methodology. Research on whether cross-examination is effective with regard to

scientific evidence is mixed, sometimes finding no effect (e.g., Kovera, McAuliff, & Hebert, 1999) or an effect for certain types of cross-examination (e.g., Austin & Kovera, 2015; Liberman et al., 2008) or finding that the efficacy of cross-examination depends on other individual (Scurich, 2015) or case-related factors (Thompson & Scurich, 2019).

Whether cross-examination might affect the present results can only be determined by additional empirical testing. Study 2 was designed to address this limitation. It is worth noting that one judge has expressed strong skepticism that cross-examination would have its intended effect in the context of firearm examiner testimony:

> [T]he Court strongly disagrees with the government that cross-examination could cure any reliability issues created by a source attribution statement . . . this discipline and the disputes surrounding it seem far too complex for a series of questions on cross-examination to allow a full understanding of the limitations of the field. . . . It would be fanciful to conclude that the normal adversarial process would enable a lay jury to adequately understand these issues, and it is similarly unrealistic to conclude that the average attorney in the average trial would be able to raise these issues in front of the jury in this fashion, particularly when this issue would be one among many issues to be presented to the jury in a trial. (*United States v. Tibbs,* 2019, pp. 52–53)

Study 2

There were two primary objectives of Study 2. The first objective was to attempt to replicate the findings in Study 1, in particular the finding that an identification increased guilty verdicts (compared to an inconclusive result) and the cannot exclude language reduced guilty verdicts compared to an identification. Replication in psychology is exceedingly important (see Open Science Collaboration, 2015).

The second objective was to enhance the ecological validity of the experiment by including a cross-examination condition (see Koehler & Meixner, 2017). Beyond ecological validity concerns, a cross-examination condition was important to include because it could potentially explain the null results for language variation observed in Study 1: If participants were predisposed to believe firearm examination is valid science, and no challenge to the testimony is presented, then it would not be surprising that variations to the language used to express the conclusion had no effect because there was no evidence presented that might alter their preconception. Alternatively, it is also possible that cross-examination, even if presented, would not affect perceptions of firearm examiner testimony. As the judge in *Tibbs* speculated, firearm examiner testimony may be too engrained in the psyche of the lay public and otherwise too technical to be adequately critiqued via cross-examination to affect verdicts. Whether cross-examination that challenges the firearm examiner has any impact alone or as a function of the conclusion language used by the expert is an empirical question that was tested in this study.

Study 2 registered the following hypotheses:

Hypothesis 1: The findings from Study 1 would replicate in that the identification conclusion increased guilty verdicts and likelihood of commission ratings relative to the cannot be excluded and inconclusive conditions.

Hypothesis 2: Cross-examination would lower guilty verdicts and likelihood of commission ratings.

Hypothesis 3: An interaction between conclusion condition and cross-examination, such that cross-examination would have no effect on guilty verdicts and likelihood of commission ratings when the conclusion is Inconclusive, because those decisions/ratings are already low, but cross-examination would significantly decrease guilty verdicts/likelihood of commission ratings for the identification and cannot be excluded conditions.

Method

Participants. Consistent with Study 1 and the related statistical power analysis, 200 participants per cell were required, for a total $N = 1260$. Participants were recruited using Qualtrics Panels and completed the survey online. Participants were terminated from the study and excluded from analyses if they did not pass a comprehension check and an attention check, if they were not a U.S. citizen, or if they came from suspicious, duplicate geolocations (see Oppenheimer et al., 2009). The final sample was comprised of 1,260 participants aged 18–92 ($Mdn = 46$, interquartile range = 29). The study was balanced with respect to gender (48% self-identified as male; 52% as female). Participants self-identified as 13.2% Black, 5.3% Asian, 62.6% White, 17.1% Hispanic, 0.2% Native American or Pacific Islander, and the rest selected other. With respect to education, 58.5% had a 2-year college degree or less, 27.6% had at least 4-year college degree, and 13.9% had a postgraduate degree.

Participants were asked to self-identify their political preferences: 13.0% identified as "very conservative," 18.0% identified as "somewhat conservative," 38.4% identified as "middle of the road," 18.1% identified as "somewhat liberal," and 12.5% identified as "very liberal." Twenty-nine percent had a self-reported annual household income of less than $20,000 and 16.9% had an income above $100,000.

Materials. The same case materials from Study 1 were used in Study 2 but with a couple of important modifications. First, there were only three variations to the language used to describe the match in Study 2:

1. *Inconclusive:* The examiner testified that "there was not sufficient agreement among the characteristics to determine that the toolmarks have been produced by the same tool."

2. *Cannot be excluded:* The examiner testified "the defendant's gun cannot be excluded as the gun that fired the bullet recovered at the crime scene," which means that "we observe some agreement of a combination of individual characteristics suggesting that the toolmarks have been produced by the same tool."

3. *Simple identification:* The examiner testified that "I identified the crime scene bullet as having been fired by the defendant's gun."

Second, some participants (randomly determined) heard cross-examination of the firearm examiner. This cross-examination was culled from transcripts of actual testimony given during cross-examination. The cross-examination presented to participants (where Q represents questions from the defense attorney and A represents answers from the firearm examiner) was as follows:

Q: You examined the crime scene bullet and test-fired bullets from the defendant's gun, correct?

A: Yes.

Q: But that required some interpretation, didn't it?

A: Well, yes, there is always an element of interpretation.

Q: It was not a perfect match, was it?

A: There were clear points of comparison on which there was consistency.

Q: But there were differences between the two bullets when you compared them under the microscope, right?

A: There are always some minor differences when you compare bullets under a microscope, but in my judgment the differences were not meaningful differences.

Q: Just to be clear, you are saying that there are differences—individual differences—between the crime scene bullet and the bullets test fired through the defendant's gun?

A: Yes, that is correct.

Q: But you do not consider those differences to be "meaningful," in your words?

A: As I said, there are always some minor differences when you compare bullets under a microscope. But a number of the individual characteristics were identical.

Q: There aren't any formal rules in your field about what constitutes a match? You just rely on your own personal judgment?

A: There is no formula for what constitutes a match because each case is unique. I rely on my knowledge, training and experience in the field to make the right judgment.

Q: So It's a match because you say it's a match?

A: Yes. I'm a certified expert in this field.

Q: Have you ever committed an error in your work?

A: No. I have passed all of my proficiency tests with perfect accuracy. Never made an error.

Q: Have you ever read a scientific study that reports an error rate for firearm examiners?

A: Maybe. I recall reading some of those studies. I believe those studies find an error rate between 0 and 1%, but as I said, my error rate is 0%.

Q: Are you aware of the 2016 report by the President's Council of Advisors on Science and Technology, which said that firearm analysis is not scientifically valid?

A: No. I have not read that report.

Q: Are you aware that mainstream scientists do not consider firearm examination to be science?

A: No. I am not aware of that and I do not agree with it.

Q: Nothing further at this time.

Thus, Study 2 used a 3 (Language: inconclusive, cannot exclude, or simple identification) × 2 (Cross-Examination: present or absent) between-participants factorial design. Participants were randomly assigned to one of six possible cells.

After reading the cross-examination transcript, participants were provided with judicial instruction on the beyond a reasonable doubt standard of proof and asked whether they would convict the defendant (dichotomous "guilty" or "not guilty"), and the likelihood that the defendant was the man who fired the gun (0–100% likelihood). Participants then responded to six different items that probed the credibility and reliability of the firearm analysis presented in the case and in general. A scale analysis revealed a high degree of correlation among the six items, yielding a Cronbach's $\alpha = .854$. Thus, a composite score (referred to as "scientific credibility") was created by summing the scores and dividing by 6. Finally, participants were asked general questions about their views about firearms evidence ("How often do firearm examiners make mistakes?" 1 = *never* to 7 = *always*; "Do you think firearm evidence is generally reliable[scientific]?" 1 = *strongly agree* to 7 = *strongly disagree*; "Do you think guns leave unique markings on discharged bullets/casings?" *yes*/*I do not know*/*no*), as well as whether erroneous convictions or failing to convict a guilty person cause more harm to society. All of the study materials, as well as underlying data, are available on an Open Science Framework website, https://osf.io/gkjts/.

Procedure. The same procedure was followed in Study 2 as in Study 1. After providing consent and responding to eligibility and background questions, participants read the crime scenario describing a simple criminal offense and police investigation and then the trial transcripts. Participants then completed the dependent measures. Average completion time was 904.72 s ($SD = 1465.30$).

Results

At the conclusion of trial, 41.9% ($n = 528$) of participants voted to convict the defendant. The percentage of guilty verdicts within each experimental condition appears in Figure 2.

A binary logistic regression model was tested with conclusion (inconclusive condition as the reference group), cross-examination (no cross-examination as the reference group), and their interaction as predictors and guilty verdict as the dependent measure. The overall model was significant ($\chi^2 = 105.59$, $df = 5$, $p < .001$, Nagelkerke $R^2 = 0.11$). See Table 2 for regression output. There was a significant main effect for conclusion ($p < .001$), with participants in both the cannot exclude condition (Exp[B] = 1.90, 95% CI [1.24, 2.91], $p < .001$) and identification condition (Exp[B] = 4.65, 95% CI [3.05, 7.08], $p < .001$) being significantly more likely to vote to convict than those in the Inconclusive condition. Moreover, participants were significantly less likely to vote to convict the defendant in the cannot exclude condition compared to the identification condition (Exp[B] = 0.41, 95% CI [0.28, 0.61], $p < .001$). These findings supported Hypothesis 1 and replicated the two findings from Study 1, which are that the identification condition produced significantly more guilty verdicts than the Inconclusive condition and the cannot exclude condition produced significantly less guilty verdicts than the identification condition. The main effect for cross-examination, how-

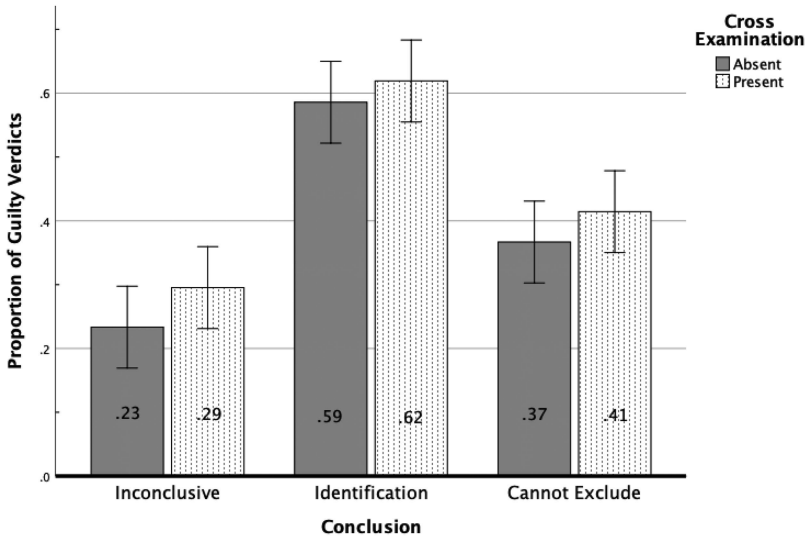

Figure 2. Proportion of guilty verdicts (with 95% confidence intervals) in each experimental condition.

ever, was not statistically significant effect ($p = .151$), nor was the interaction term ($p = .831$).

Participants' ratings of the likelihood that the defendant was the man who fired the gun were examined next. A 3(Conclusion) × 2(Cross-Examination) ANOVA revealed a significant main effect for conclusion, $F(2, 1254) = 64.77$, $p < .001$, $\eta_p^2 = 0.094$ such that the identification conclusion yielded the highest likelihood ratings ($M = 68.89$, $SE = 1.35$), followed by cannot be excluded ($M = 59.73$, $SE = 1.35$), and then inconclusive ($M = 47.181$, $SE = 1.35$). This main effect is consistent with the main effect detected for guilty verdicts. Post hoc comparisons with Bonferroni adjustments revealed that each of these marginal means were significantly different from each other at the $p < .001$ level, with identification yielding larger likelihood estimates than the Inconclusive wording (Cohen's $d = 0.77$) and cannot be excluded ($d = 0.34$), which in turn was significantly higher than inconclusive ($d = .44$). Again, this supported Hypothesis 1 and replicated findings from Study 1.

Similar to the pattern observed with guilty verdicts, the main effect for cross-examination was not statistically significant, $F(1, 1254) = 3.16$, $p = .076$, $\eta_p^2 = 0.003$. However, in contrast to the pattern of guilty verdicts, the interaction between cross-examination and conclusion was statistically significant, $F(2, 1254) = 3.97$, $p = .019$, $\eta_p^2 = 0.006$. The means (and 95% CIs) are plotted in Figure 3.

As is apparent in Figure 3, when there is no cross-examination, there is a difference between each of the three conclusion conditions; however, when cross-examination is present there is a difference between the inconclusive conclusion but not the identification or cannot exclude conditions. In other words, cannot exclude and identification are perceived as different conclusions without cross-examination but they are indistinguishable when cross-examination is present. Although this pattern of results is statistically significant, it is important to point out that the effect is quite small in practical terms, a point we return in the discussion section.

A 3 × 2 ANOVA with scientific credibility as the dependent measure detected a main effect for conclusion, $F(2, 1254) = 7.83$, $p < .001$, $\eta_p^2 = 0.012$, and a main effect for cross-examination, $F(1, 1254) = 33.59$, $p < .001$, $\eta_p^2 = 0.026$, but the interaction was not statistically significant, $F(2, 1254) = 2.48$, $p = .084$, $\eta_p^2 = 0.004$. The mean scientific credibility values for inconclusive, cannot exclude, and identification are 5.52 ($SD = .97$), 5.56 ($SD = .92$), 5.77 ($SD = 1.04$), respectively, and both inconclusive and cannot exclude were significantly less scientifically credible than

Table 2
Study 2 Binary Logistic Regression Predicting Verdict (1 = Guilty, 0 = Not Guilty)

Experimental condition	Exp(*B*) (*SE*)	95% CI for Exp(*B*) Lower	Upper
Constant	0.34 [.001] (0.163)		
Simple identification	4.65 [.001] (0.22)	3.05	7.08
Cannot be excluded	1.90 [.003] (0.22)	1.24	2.91
Cross-examination	1.38 [.15] (0.22)	0.89	2.13
Simple Identification × Cross Examination	0.84 [.55] (0.30)	0.47	1.50
Cannot Be Excluded × Cross Examination	0.89 [.69] (0.30)	0.49	1.60
N	1,260		
Nagelkerke R^2	0.11		

Note. CI = confidence interval. *p* Values in brackets.

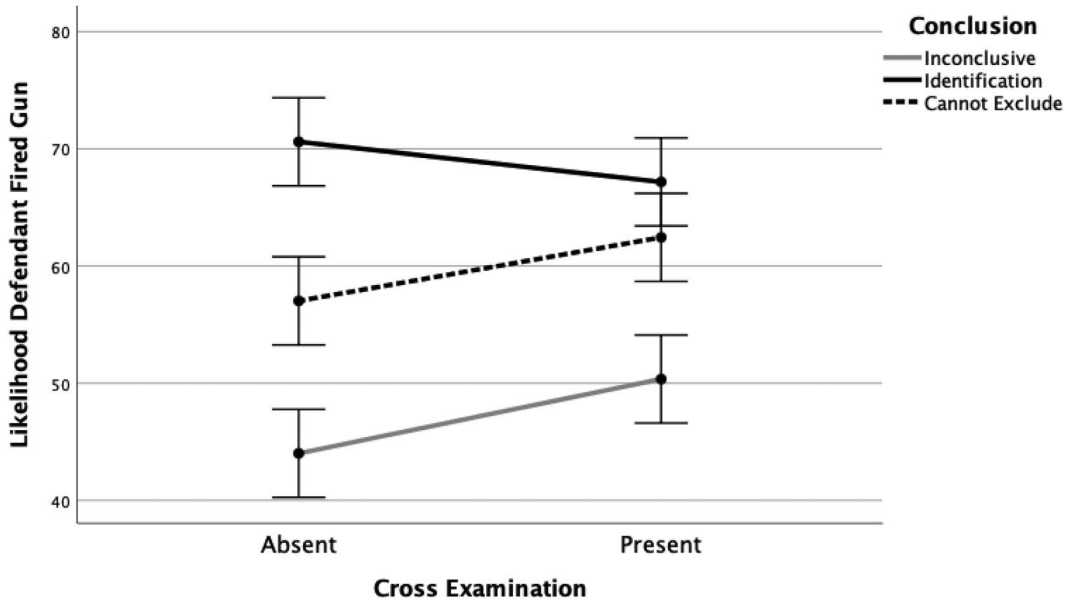

Figure 3. Mean likelihood ratings that the defendant fired the gun (with 95% confidence intervals) in each experimental condition.

identification ($p_{\text{bonferroni}} < .001$, and $p_{\text{bonferroni}} = .004$, respectively). When cross-examination is absent, the mean scientific credibility rating is 5.77 ($SD = 0.90$) whereas the mean scientific credibility decreased (5.46 [$SD = 1.03$]) when cross-examination was present. Again, this difference is statistically significant (as evidenced by the main effect reported above), but it is quite meager in terms of its practical significance.

Participants' general attitudes toward firearms forensic evidence were also measured. Consistent with Study 1, participants, on average, believe firearm experts rarely make mistakes in their analysis very often ($M = 2.59$, $SD = 0.97$, between *almost never* and *sometimes* anchors), and that firearm evidence in general is reliable ($M = 2.50$, $SD = 1.34$, between *agree* and *somewhat agree*) and based on scientific principles ($M = 2.47$, $SD = 1.26$, between *agree* and *somewhat agree*). Finally, 83.8% of participants ($n = 1{,}056$) stated that they believe firearms leave unique markings on bullet casings; 13.7% ($n = 173$) said that they were unsure, and 2.5% ($n = 31$) said firearms do not leave unique markings. These numbers are nearly identical to responses from Study 1, suggesting that they are fairly robust.

General Discussion

These results can be viewed with a mix of optimism and pessimism. On the one hand, forensic experts can use wording that more cautiously describes their conclusions without harming their credibility or conviction rate. Thus, there is little to no cost in avoiding misleading or overstated conclusion language that implies a categorical association between evidence. On the other hand, it is unfortunate that the more cautious and tempered conclusion language did not result in higher credibility ratings for the expert or reliability and scientific ratings for the evidence.

In general, participants seemed to hold firearm forensic analysis in high regard, believing errors do not happen very often, and agreeing that it is scientific and reliable. In fact, the vast majority of participants believe that firearms make unique marks on bullets and casings. As a result, those participants may assume that any conclusion reached, however it is phrased, may involve a unique and infallible "match."

The modifications to forensic firearms analysis conclusions that have been adopted by federal judges and the DOJ may be ineffective because the modifications do not address the error rates and the limitations of firearms analysis, in a way that could counter participants' prior assumption that firearms conclusions are largely infallible and reach some sort of "perfect match." Using the phrase "reasonable degree of ballistic certainty" is itself unclear; it has no technical meaning and nor is it common in everyday usage. In contrast, the phrase "more likely than not" did convey an actual probability, but not one that an expert can actually explain or defend because it is not supported by any empirical data.

Another approach is to provide jurors with more information about the reliability of the method. Presenting information about empirical evidence concerning error rates, proficiency of firearms examiners (Mitchell & Garrett, 2019), or presenting information in a quantitative form based on actual population data (Garrett et al., 2018), might all better improve the accuracy of jury decision making. Jurors are highly receptive to such information (Mitchell & Garrett, 2019). To be sure, research suggests that jurors can sometimes struggle to understand and amalgamate error rate information, particularly if it is not accompanied by an account of how forensic science errors can occur (see Scurich, 2015).

Cross-examination did not help jurors to consistently discount firearms conclusions, consistent with prior work showing mixed effects of cross-examination on jury perceptions of strength of evidence. Indeed, the cannot exclude and identification conclusions were indistinguishable when cross-examination was present. One reason may be that, as one might expect, during cross-

examination the expert did not disavow the work, and instead continued to maintain that, based on training and experience, the same conclusion should be reached. The cross-examination permitted the defense counsel to highlight certain limitations of the firearms comparison method, but the expert largely stuck to his guns, as it were. The results suggest that the judge in *Tibbs* may have been right to be skeptical that cross-examination is the solution to the problem of overstated firearms testimony (*United States v. Tibbs,* 2019, pp. 52–53).

A challenge for a discipline like firearms comparison, is that the error rate of firearms examiners is not sufficiently well understood. As one judge noted,

> after extensive review of the testimony of the expert witnesses and of the studies about which those experts testified, the undersigned finds it difficult to conclude that the existing studies provide a sufficient basis to accept the low error rates for the discipline that these studies purport to establish. (*United States v. Tibbs,* 2019, p. 28)

Further basic research must be conducted to measure error rates in the firearms discipline, and such work is apparently underway (National Institute of Standards and Technology, 2020). Similarly, as with any other type of forensic pattern comparison analysis, the method can be subject to error due to cognitive bias (Kassin, Dror, & Kukucka, 2013). However, studies have not been done examining the effects of task irrelevant information or other biasing information on firearms examiners. Nor have studies been done regarding whether reliability of firearms evidence varies given different levels in evidence quality and difficulty (Dror & Scurich, in press).

Studies have found that individual juror perceptions of the strength of the prosecution case strongly predict their guilt verdicts (Bornstein & Green, 2011). Across conclusions that indicate a match, approximately 60% of participants in each condition believed the single piece of firearms comparison evidence met the reasonable doubt threshold to convict the defendant. Like the participants in this study, lay jurors in a criminal case have no reason not to assume that a firearms expert makes a highly reliable match based on "unique" markings. Nor can existing research provide them more calibrated information, because as noted, the strengths and the weaknesses of the methods that firearms experts use have not been adequately tested empirically. Jurors can vary the weight that they place on forensic evidence that they would otherwise treat as extremely probative, such as fingerprint evidence, if they receive quantitative information regarding proficiency, or the reliability of a particular forensic examiner (Mitchell & Garrett, 2019) or regarding a forensic conclusion (Garrett et al., 2018). Perhaps judges or lawyers can similarly use such quantitative information to inform jurors regarding firearms evidence. Doing so, however, will require better underlying research on the basic reliability and validity of firearms comparison methods.

Limitations and Future Directions

The current studies have several limitations that need to be taken into account. The stimulus materials presented an abbreviated version of a criminal trial in written format. This deviates from an actual criminal trial in many obvious ways. For instance, in an actual trial, jurors would watch live testimony from a firearm examiner and likely be furnished with images of the ammunition in question. A nontrivial amount of time would be spent in the expert's qualifications and training. Jurors would deliberate and reach a verdict, rather than offer individual guilty verdicts. It is not clear how these differences might impact the observed results. Future research should include video testimony so to better mimic the real-world condition in which jurors view the expert testifying (e.g., Scurich, 2018), and include more information regarding the expert's background and training (Koehler, Schweitzer, Saks, & McQuiston, 2016). Generalizing the present findings to different cases and different contexts should be done cautiously.

In addition to enhancing ecological validity, future research might examine if jurors are sensitive to the differences between class, subclass, and individual characteristics. According to the AFTE theory, a "match" can only be declared when there is sufficient agreement of individual characteristics. But it is certainly possible that jurors assume that agreement of class or subclass characteristics is sufficient to declare a "match." This speculation is supported in part because the stimulus materials in Study 1 (Conditions 6 and 7) inappropriately said that an identification was reached on the basis of class characteristics and this apparently did not affect verdicts. A more careful empirical test of this possibility is necessary, and particularly important as courts start to consider limiting firearm examiner testimony to class or subclass characteristics only.

Conclusion

While several federal judges and the DOJ have tempered language that firearms experts use in court, adopting types of modified conclusion language endorsed by practitioners did not affect guilty verdicts. In contrast, a conclusion that an examiner simply "cannot exclude the defendant's gun," a much more cautious conclusion, imposed by at least one judge, did affect guilty verdicts. These findings suggest that many judicial and prosecution-driven interventions to limit conclusion language for firearms testimony are not likely to be effective. Any courtroom interventions must be quite forceful to be effective, because as the studies here revealed, laypeople place great weight on firearms testimony. Judicial actors require adequate research to assess the reliability of firearms comparison methods, to use that research to more carefully inform jurors in criminal cases. Until that foundational research is conducted, however, more forceful judicial instructions may be warranted.

References

Association of Firearms and Tool Mark Examiners. (1998). Theory of identification as it relates to toolmarks. *AFTE Journal, 30,* 86–98.

Association of Firearms and Tool Mark Examiners. (2020). *Theory of identification as it relates to toolmarks*. Retrieved from https://afte.org/about-us/what-is-afte/afte-theory-of-identification

Austin, J. L., & Kovera, M. B. (2015). Cross-examination educates jurors about missing control groups in scientific evidence. *Psychology, Public Policy, and Law, 21,* 252–264. http://dx.doi.org/10.1037/law0000049

Bornstein, B. H., & Green, E. (2011). Jury decision making: Implications for and from psychology. *Current Directions in Psychological Science, 20,* 63–67. http://dx.doi.org/10.1177/0963721410397282

Bureau of Justice Statistics. (2011). *Nonfatal firearm violence, 1993–2011, special tabulation from the Bureau of Justice Statistics' National Crime Victimization Survey*. Retrieved from https://www.nij.gov/topics/crime/gun-violence/pages/welcome.aspx

Dror, I. E., & Scurich, N. (in press). (Mis)use of scientific measurements in forensic science. *Forensic Science International: Synergy*.

Federal Bureau of Investigation. (2018). *Crime in the United States*. Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2018/preliminary-report

Garrett, B. L., Crozier, W. E., & Grady, R. (2020). Error rates, likelihood ratios, and jury evaluation of forensic evidence. *Journal of Forensic Sciences, 65,* 1199–1209. http://dx.doi.org/10.1111/1556-4029.14323

Garrett, B., & Mitchell, G. (2013). How jurors evaluate fingerprint evidence: The relative importance of match language, method information and error acknowledgement. *Journal of Empirical Legal Studies, 10,* 484–511. http://dx.doi.org/10.1111/jels.12017

Garrett, B., Mitchell, G., & Scurich, N. (2018). Comparing categorical and probabilistic fingerprint evidence. *Journal of Forensic Sciences, 63,* 1712–1717. http://dx.doi.org/10.1111/1556-4029.13797

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory & Cognition, 2,* 42–52. http://dx.doi.org/10.1016/j.jarmac.2013.01.001

Kennedy, D. (2003). Forensic science: Oxymoron? *Science, 302,* 1625. http://dx.doi.org/10.1126/science.302.5651.1625

Koehler, J. J., & Meixner, J. B. (2017). Jury simulation goals. In M. B. Kovera (Ed.), *The psychology of juries* (pp. 161–183). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/0000026-008

Koehler, J., Schweitzer, N. J., Saks, M., & McQuiston, D. (2016). Science, technology or the expert witness: What influences judgments about forensic science testimony? *Psychology, Public Policy, and Law, 22,* 401–413. http://dx.doi.org/10.1037/law0000103

Kovera, M. B., McAuliff, B. D., & Hebert, K. S. (1999). Reasoning about scientific evidence: Effects of juror gender and evidence quality on juror decisions in a hostile work environment case. *Journal of Applied Psychology, 84,* 362–375. http://dx.doi.org/10.1037/0021-9010.84.3.362

Lanigan, B. (2012). Firearms identification: The need for a critical approach to, and possible guidelines for, the admissibility of "ballistics" evidence. *Suffolk Journal of Trial & Appellate Advocacy, 17,* 54.

Lieberman, J. D., Carrell, C. A., Miethe, T. D., & Krauss, D. A. (2008). Gold versus platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychology, Public Policy, and Law, 14,* 27–62. http://dx.doi.org/10.1037/1076-8971.14.1.27

McQuiston-Surrett, D., & Saks, M. J. (2009). The testimony of forensic identification science: What expert witnesses say and what factfinders hear. *Law and Human Behavior, 33,* 436–453. http://dx.doi.org/10.1007/s10979-008-9169-1

Mitchell, G., & Garrett, B. L. (2019). The impact of proficiency testing information and error aversions on the weight given to fingerprint evidence. *Behavioral Sciences & the Law, 37,* 195–210. http://dx.doi.org/10.1002/bsl.2402

Mnookin, J. L., Cole, S. A., Dror, I. E., & Fisher, B. A. (2010). The need for a research culture in the forensic sciences. *UCLA Law Review, 58,* 725–780.

National Institute of Standards and Technology. (2020). *NIST and NOBLIS seek participants for Bullet Black Box Study*. Retrieved from https://forensicstats.org/blog/2020/04/06/nist-and-noblis-seek-participants-for-bullet-black-box-study/

National Research Council. (2008). *Ballistic imaging*. Washington, DC: The National Academies Press.

National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: The National Academies Press.

Open Science Collaboration. (2015). Psychology: Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872. http://dx.doi.org/10.1016/j.jesp.2009.03.009

President's Council of Advisors on Science and Technology. (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods*. Retrieved from http://www.crime-scene-investigator.net/PDF/forensic-science-in-criminal-courts-ensuring-scientific-validity-of-feature-comparison-methods.pdf

Ribeiro, G., Tangen, J. M., & McKimmie, B. M. (2019). Beliefs about error rates and human judgment in forensic science. *Forensic Science International, 297,* 138–147. http://dx.doi.org/10.1016/j.forsciint.2019.01.034

Saks, M. J., & Wissler, R. L. (1984). Legal and psychological bases of expert testimony: Surveys of the law and of jurors. *Behavioral Sciences & the Law, 2,* 435–449. http://dx.doi.org/10.1002/bsl.2370020410

Schklar, J., & Diamond, S. S. (1999). Juror reactions to DNA evidence: Errors and expectancies. *Law and Human Behavior, 23,* 159–184. http://dx.doi.org/10.1023/A:1022368801333

Scurich, N. (2015). The differential effect of numeracy and anecdotes on the perceived fallibility of forensic science. *Psychiatry, Psychology and Law, 22,* 616–623. http://dx.doi.org/10.1080/13218719.2014.965293

Scurich, N. (2018). What do experimental simulations tell us about the effect of neuro/genetic evidence on jurors? *Journal of Law and the Biosciences, 5,* 204–207. http://dx.doi.org/10.1093/jlb/lsy006

Scurich, N., & John, R. S. (2013). Mock jurors' use of error rates in DNA database trawls. *Law and Human Behavior, 37,* 424–431. http://dx.doi.org/10.1037/lhb0000046

Thompson, W. C., Kaasa, S. O., & Peterson, T. (2013). Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies, 10,* 359–397. http://dx.doi.org/10.1111/jels.12013

Thompson, W. C., & Newman, E. J. (2015). Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior, 39,* 332–349. http://dx.doi.org/10.1037/lhb0000134

Thompson, W. C., & Scurich, N. (2019). How cross-examination on subjectivity and bias affects jurors' evaluations of forensic science evidence. *Journal of Forensic Sciences, 64,* 1379–1388. http://dx.doi.org/10.1111/1556-4029.14031

Tyler, T. R. (2005). Viewing CSI and the threshold of guilt: Managing truth and justice in reality and fiction. *The Yale Law Journal, 115,* 1050–1085. http://dx.doi.org/10.2307/20455645

United States v. Diaz, No. CR 05–00167 WHA, 2007 WL 485967(N. D. Cal. Feb. 12, 2007).

United States v. Driscoll, 1:05-cr-00100-RWR (E. D. Pa. Feb. 10, 2003).

United States v. Glynn, 578 F. Supp. 2d 567 (S. D. N. Y. 2008).

United States v. Monteiro, 407 F. Supp. 2d 351 (D. Mass. 2006).

United States v. Tibbs, 2016 CF1 019431 (Sup. Ct. D. C. 2019).

United States v. Willock, 696 F. Supp. 2d 536 (D. Md. 2010).

U.S. Department of Justice. (2019). *Approved ULTR for the Forensic Firearms/Toolmarks Discipline* [Guideline memorandum]. Retrieved from https://www.justice.gov/olp/page/file/1083671/download

Received June 9, 2020
Revision received August 4, 2020
Accepted August 10, 2020 ■